DOCUMENT RESUME

ED 022 173 AL 001 471

By-Baehr, Timothy J.

TOWARD THE QUANTITATIVE ANALYSIS OF 'DEVIANT' ARTICULATION.

Michigan Univ., Ann Arbor. Dept. of Psychology.
Spons Agency-National Inst. of Child Health and Human Development, Bethesda, Md.
Report No-R-35
Pub Date 10 Jul 67
Note-27p.
EDRS Price MF-$0.25 HC-$1.16
Descriptors-APHASIA, *ARTICULATION (SPEECH), CHILD LANGUAGE, CONTRASTIVE LINGUISTICS, *DATA ANALYSIS, DISTINCTIVE FEATURES, PHONEMICS, *PHONETIC ANALYSIS, *PHONETIC TRANSCRIPTION, SECOND LANGUAGE LEARNING, *SPEECH EVALUATION, SPEECH HANDICAPPED
Identifiers-Analphabetic transcription, *Theory of Signal Detectability

The evaluation of 'deviant articulation' (that of young children, speech defective persons, aphasics, second-language learners) has usually consisted of two activities: transcription of the speech being evaluated, and comparison of the transcription against some 'standard' set of 'target' sounds. Any transcription is a description of a speaker's articulation in terms of the auditory and perceptual capacities of a transcriber. This paper presents a method for the transcription and quantitative analysis of deviant articulation. The method is based on (1) analphabetic transcription using binary articulatory categories , and (2) analysis of the transcripion in terms of the perceptual performance of the transcriber as measured by the Theory of Signal Detectability (Tanner and Birdsall, 1958). This paper comprises a report in "Development of Language Functions, A Research Program-Project (Study C: The Development of Speech and Sound Specificity in Children)." (DO)

The University of Michigan

Department of Psychology

Toward the Quantitative Analysis of 'Deviant' Articulation

by

Timothy J. Baehr

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Report Number 35

Development of Language Functions

A Research Program-Project

(Study C: The Development of Speech and Sound Specificity in Children)

July 10, 1967

Supported by the National Insitute for Child

Health and Human Development

Grant Number 5 PO1 HD 01368-03

ED022173

AL 001 471

Toward the Quantitative Analysis of 'Deviant' Articulation

Timothy J. Baehr

University of Michigan

The evaluation of 'deviant' articulation (e.g. that of young children, speech defective persons, aphasics, second-language learners) has usually consisted of two activities: transcription of the speech being evaluated, and comparison of the transcription against some 'standard' set of 'target' sounds.

Any transcription is a description of a speaker's articulation in terms of the auditory and perceptual capacities of a transcriber. Obviously, a poorly trained or inattentive transcriber will produce false or misleading descriptions of speakers' articulation. However, even well trained linguists must have different hearing sensitivities and different criteria for assigning various symbols to speech sounds.

The aim of this paper is to present a method for the transcription and quantitative analysis of deviant articulation. The perceptual abilities of the transcriber are accounted for and made an integral part of the method.

Two approaches to the evaluation of articulation correspondences

The simplest (or at least the most widespread) strategy in making a transcription and subsequent comparison of deviant articulation against some standard is to express both the speaker's output and the standard in terms of a phonetic alphabet. A qualitative comparison can be made, and percents correct can be calculated for each of the target sounds. This 'alphabetic' approach has several shortcomings: (1) Nothing is said about the articulatory positions or processes involved, (2) The process the transcriber goes through in order to decide which of several hundred alphabetic symbols to select for a given sound is not overtly indicated in his transcription, (3) Quantifications (such as percent correct) have questionable meaningfulness.

Most investigators rarely stop, however, at a simple alphabetic transcription. They re-analyze, at least partly, their alphabetic data in terms of some phonetic attributes.

Re-analysis doesn't add anything new to the data; it reorganizes it into a more interesting or more revealing form. Thus, an 'analphabetic' approach is usually a tautological, *a posteriori* re-analysis of an alphabetic transcription.

On the other hand, there should be no theoretical obstacle to starting *a priori* with a suitable analphabetic transcription method based on articulatory parameters. The analphabetic method offers an experimenter the opportunity to determine how successfully transcribers can use articulatory parameters as perceptual categories in making an analphabetic transcription.

Analphabetic classification system

General principles: Speech is segmentable into descrete entities which may be called 'phones'. Just how segmentation is to take place is a problem which has plagued linguists and designers of speech-recognition devices for years. Although almost any linguist has little trouble segmenting the speech he wishes to transcribe and analyze, the process of segmentation has eluded exact specification. Some elucidation of the processes and principles of segmentation may be found in Pike (1943). For the purposes of this paper, it will be assumed that segmentation is a perceptual process distinct from description, and that all transcribers have equally good skills of segmentation.

Since no two phones ever uttered will be exactly alike, it would in practice be impossible to symbolize phones unless the symbols constituted a nearly infinite set. On the other hand, phones may be classified together in a finite set of 'phone types', each phone type being assigned a different symbol. The classification of phones into phone types may be considered the basis of alphabetic transcription.

A phone type, then, is defined as the simultaneous intersection of a segment and the terms of a classificational system. The classificational terms constitute the analphabetic component of transcription. For the purpose of the present paper, it will be convenient to specify some formal and practical requirements that the classificational system and its terms must meet:

(1) The system must be composed of a finite set of categories.

(2) The categories must be relatable to observables at at least one stage (e.g. articulation, acoustic signal, audition, perception, decoding) of the speech event.

(3) The categories must yield phonetic descriptions (either in terms of the categories themselves or the phone types derived from the category com-complexes) adequate for at least distinguishing dialects from one another.

(4) The categories must be uniform, that is, each category must have the same degrees of membership or specificity.

(5) The categories may overlap somewhat, but not to the point of yielding an ambiguous or phonetically redundant description of a given phone type.

It should be obvious that there is an intimate relationship between (1), (4), and (5). The descriptive power of the categories and the resultant phone types is a function of the number of categories, their degree of specificity, and the amount of overlap among them.

Only the first requirement is strictly formal. The others are necessary for the elegance and manageability of the analphabetic system.

Articulatory categories: A set of phonemic categories that otherwise seems to fit the above requirements fairly well is Jakobson's 'distinctive features' (Jakobson, Fant, and Halle, 1963; Jakobson and Halle, 1956). Finer phonetic distinctions among phone types may be obtained if the feature system can be expanded and modified somewhat. This expansion and modification is discussed below.

Jakobson and his colleagues relate the feature system to various stages of the speech event. Their terms are a combination of articulatory (e.g. 'Nasal'), acoustic ('Diffuse'), and auditory ('Strident') terminology. However, since the transcriber would be making judgments about the articulatory performance of speakers, such a combination of terminologies would seem unnecessary and possibly disruptive, unless the transcriber were already familiar with distinctive features and their articulatory correlates. Therefore, it is assumed that the feature terminology may be easily translated into 'articulatory categories'. Such a translation would maintain the uniformly binary nature of the categories, and some of the theoretical reasons for binariness (Jakobson, Fant, and Halle, 1963, Chapter 1). The most salient point, however, especially in regard to requirement (4) above, is the uniformity of this binariness across all categories.

A tentative set of articulatory categories is described briefly below, with deviations from and additions to Jakobsonian terminology noted. Basic to much of the description are Halle's (1964) postulated four degrees of narrowing in the vocal tract: Contact, Occulusion, Obstruction, and Constriction. These four degrees of narrowing are characterized by stops, fricatives, glides, and high vowels, respectively.

VOCALIC--NON-VOCALIC. Vocalic sounds have a degree of narrowing not exceeding constriction; Non-Vocalic sounds have a degree of narrowing that exceeds constriction.

CONSONANTAL--NON-CONSONANTAL. Consonantal sounds have a degree of narrowing equal to or exceeding occlusion; Non-Consonantal sounds have a degree of narrowing less than occlusion.

INTERRUPTED--CONTINUANT. Interrupted sounds have a degree of narrowing equal to contact; Continuants have a degree of narrowing less than contact. Certain nasals (e.g. [m], [n], etc.), because they open an alternate 'escape route' for the airstream (i.e. the nasal passage), are described as Continuant.

EDGED--NON-EDGED. Edged sounds involve the forcing of the airstream over a relatively sharp edge, such as the teeth or uvula. In addition, Edged sounds must have a degree of narrowing that exceeds constriction. The Jakobsonian term for this category is Strident-Mellow.

PERIPHERAL-1--NON-PERIPHERAL-1. Peripheral-1 sounds have a primary narrowing at either of the oral peripheries, the lips or the velum; Non-Peripheral-1 sounds have their primary narrowing elsewhere. The Jakobsonian term is Grave--Non-Grave.

PERIPHERAL-2--NON-PERIPHERAL-2. Peripheral-2 sounds have a secondary narrowing at one of the oral peripheries. Non-Peripheral-2 sounds either do not have a secondary narrowing or have one located elsewhere. The Jakobsonian term is Flat-Plain.

MEDIAL-1--NON-MEDIAL-1. Medial-1 sounds are articulated with a primary narrowing in the middle of the vocal cavity; Non-Medial-1 sounds have their primary narrowing elsewhere. The Jakobsonian term is Acute--Non-Acute. This category generally applies only to the vowels and glides; for the consonants and liquids, Peripheral-1--Non-Peripheral-1 usually provides adequate descriptive specificity.

MEDIAL-2--NON-MEDIAL-2. Medial-2 sounds have a secondary narrowing at the palate; Non-Medial-2 sounds either do not have a secondary narrowing or have one elsewhere. The Jakobsonian term is Sharp-Plain.

CLOSE--NON-CLOSE. Close sounds must be articulated with the mandible closed or nearly closed. Thus, the front consonants and liquids and the high vowels and glides cannot be articulated if the mandible is opened beyond a certain point, without strenuous compensation in the pharyngeal cavity. The Non-Close sounds do not have this restriction on mandible-closure. The Jakobsonian term is Diffuse--Non-Diffuse.

OPEN--NON-OPEN. Open sounds are articulated such that the mandible <u>may</u> be opened to its widest extent. No degree of narrowing equal to or exceeding constriction is possible. Therefore, this term describes the open, or wide, vowels. Non-Open sounds cannot be articulated with the mandible opened to the extent permitted among the Open sounds. The Jakobsonian term is Compact--Non-Compact.

NASAL--NON-NASAL. Nasal sounds are those produced by directing part of all of the airstream through the nasal cavity; Non-Nasal sounds are produced by directing all of the airstream through the oral cavity.

VOICED--VOICELESS. Voiced sounds are produced with vocal cord vibration; Voiceless sounds are produced without vocal cord vibration.

TENSE--LAX. In Tense sounds, the articulators spend a relatively longer time away from a neutral 'rest' position than in Lax sounds. (An alternative definition has to do with the relative amount of air pressure posterior to the point of narrowing. The writer finds that definition less convincing.)

The next three categories are due to J. C. Catford (personal communication, 1965); the writer, however is responsible for their particular application.

EGRESSIVE--INGRESSIVE. Egressive sounds are those in which the airstream flows out of the vocal tract; Ingressive sounds are those in which the airstream flows into the vocal tract.

PULMONIC--NON-PULMONIC. In Pulmonic sounds, the airstream is initiated at the lungs (by pressure of the diaphragm and/or the intercostal muscles); in Non-Pulmonic sounds, the airstream is initiated elsewhere. The vowels are all Pulmonic.

GLOTTALIC--VELARIC. Glottalic sounds are those which are initiated by compression or expansion of the glottis; Velaric sounds are initiated at the velum. The closest corresponding Jakobsonian term is Checked--Unchecked, although its definition is somewhat different. Sounds can be described as Glottalic or Velaric only if they are also Non-Pulmonic.

Stress, tone, and length are not included in the set of articulatory categories. The writer feels that these three aspects of speech are more properly evaluated as part of the analysis of syllables, not segments. Moreover, stress, tone, and length can be made binary only in the most artificial way.

It seems at this point that articulatory facts are being set up as perceptual categories or features. In a sense, this is true; but since the transcriber shares the same articulatory mechanisms with the speaker, it is reasonable for him to perceive in terms of articulatory parameters. Moreover, if one considers the features to be merely names or labels of corresponding auditory and perceptual phenomena, then the question of terminology becomes trivial.

Phone types: A phone type is defined as the simultaneous intersection of a 'segment' with the sixteen articulatory categories. Each segment-category intersection is marked in one of three ways: With (1) a Plus or (2) a Minus, indicating whether the segment has one or the other value of the binary category, or (3) a Zero, indicating an 'impossible' intersection due to the way the categories have been defined. For example, the vowels (marked Plus Vocalic and Minus Consonantal) cannot be also marked Plus or Minus Edged. Likewise, Plus or Minus Open cannot be used to describe segments otherwise marked as Plus Consonantal.

The Standard: The standard articulations against which a given deviant speaker is to be compared are presented as a chart. Across the top of the chart is an inventory of the phone types of the speech community being used as the standard of comparison. Down

the left side of the chart are the articulatory categories. At the intersection of a given category with a given phone type a 'plus' or 'minus' indicates whether the phone types has one or the other attribute of the category. Impossible intersections are marked with a 'zero'. The impossible intersections are not language-specific; they are due to the way the categories are defined.

An illustration of what a Standard should look like is given in Table 1. The Standard is of the writer's dialect of Midwest American.

Insert Table 1 about here

Sometimes it may be desirable to introduce further zeroes into the Standard. This would be the case if there were a range of acceptable phone types that were acceptable variant articulations in a given test item. For instance, if it made no difference whether the speaker pronounced 'frog' as [frag] or [frɔg], then the vowel nucleus would be marked Ø Peripheral-2 and Ø Tense in the Standard.

Collection of speech samples from deviant speakers

The specifice procedures for collecting speech samples will vary depending on the speaker or speakers to be tested. However there are some general criteria that should be observed:

(1) If at all possible, the speech sample should be elicited in the form of single-word isolates, that is, preceded and followed by a pause. This is to reduce the possibly unpredictable effect of syntactic, morphological and prosodic environment on articulation. As an alternative, the speech may be elicited as part of a consonant verbal framework. This criterion need not apply, of course, if the verbal framework is manipulated as an independent variable.

(2) Unless imitative ability is being evaluated, the speech sample should not be obtained by imitation. Instead, the speaker may be asked to name certain items or actions, answer leading questions (such as "What is the opposite of hot?"), etc. Imitation would, of course, allow the use of nonsense syllables if this were required in the experimental design.

(3) Each word (or nonsense syllable) elicited from the speaker must be identified; otherwise the transcriber will have no idea what the target sounds were. One of the easiest ways to obtain identification is to have a specified order of presentation of the stimuli. An even safer strategy is to include the identifications on the sample tape recordings.

(4) The sample size should be large enough to allow for statistical analysis of the transcription(s). There can be no rigid rule for sample size, but the sample should probably include not less than 75-100 words (containing 375-500 elicited phones).

(5) The sample should be recorded on the best recording equipment available.

Transcription method

Straight transcription: Two tape recorders are used. One plays the tape recording obtained from the speaker; the other is fitted with a recursive tape loop of about eight seconds' duration at 7 1/2 i.p.s. The tape machine plays into the loop machine, which is on 'Record' mode. The transcriber listens to the tape through earphones. When he hears a word spoken by the speaker, he switches the tape machine to 'Stop' and the loop machine to 'Play'. This allows the transcriber to listen to the utterance repeatedly without continually back-spacing the main tape. Mode-switching on the tape recorders is by remote control. See Figure 1.

Insert Figure 1 about here

The transcriber uses a transcription form similar to the one in Figure 2. As he listens to an utterance, the transcriber marks each segment-category intersection with a plus or minus to indicate what he has heard. Thus, for each speech sound the transcriber hears, he has to make sixteen decisions about the parameters of the sound. Impossible or other Zero intersections (as described above) are either left blank or marked with a Zero.

Insert Figure 2 about here

Differential transcription: Differential transcription differs from straight transcription in that the transcriber marks on the transcription form only those segment-category intersections which were 'missed' by the speaker, vis à vis the Standard. Although the transcriber still must make sixteen decisions about each speech sound, the tedium of having to mark every intersection on the form is eliminated. The 'correct' intersections

can be recovered later, if necessary, by referring back to what the target sounds were in terms of the standard.

Training: Training consists of two steps: (1) Familiarization with the articulatory categories and their definitions, and (2) Transcription of a tape of stimulus-words or nonsense syllables in which the articulatory parameters are known and controlled by the speaker. After each stimulus, the correct transcription is given and the trainee's mistakes are discussed.

Perception and decision processes in transcription

A transcription method that accounts for the perceptual abilities of the transcriber must include an explicit statement of its underlying theory of perception. The material in this and the following section is based almost entirely on the Theory of Signal Detectability (Tanner and Birdsall, 1958; Clarke, Birdsall, and Tanner, 1959; Swets, Tanner and Birdsall, 1961; and others, cf. Swets, 1964).

Definition of 'Signal': A 'Signal' will be defined as that part of the speaker's acoustic output which carries information about the articulatory positions and processes used in speaking. This acoustic output need not be distinct from or independent of that which carries other information such as rhythm, stress, syntax, affective state of the speaker, etc. An 'Ideal Signal' is defined as that which is produced by a normal speaker using his native language or a trained speaker producing nonsense syllables (cf. a previous paragraph on Training). A 'Non-Ideal Signal' is that which is produced by a (real or suspected) deviant speaker.

For each target sound attempted (by either a normal or deviant speaker) there are about sixteen 'intentions' attributable to the speaker concerning the articulatory components of the sound. These sixteen intentions correspond to the articulatory categories which simultaneously intersect the target sound segment (no intentions are attributable for impossible or Zero intersections). Clearly, intention is a hypothetical construct and does not imply that the speaker is usually aware of manipulating the various articulatory parameters.

Because of the way the articulatory categories are defined, each intention is binary. However, the physical realization of the intention is not binary. For example, if in a given sample of speech, 100 sounds were intended to be voiceless, they would not all be absolutely voiceless in their realization. Because of other factors (articulatory positions, prosody, etc.), some intended voiceless sounds may be very slightly voiced. The distribution of the 100 intended voiceless sounds might look like Fig. 3.

Insert Figure 3 about here

In the same sample of speech, 100 sounds may have been intended to be voiced. Again, the physical realization of this intention will exhibit a certain distribution along the degree of voicing continuum. This distribution of intended voiced sounds may be plotted on the same coordinates as the intended voiceless sounds, as in Fig. 4.

Insert Figure 4 about here

It will be noticed that in Fig. 4 the two distributions overlap slightly. This overlap, to a greater or lesser extent, is to be expected for all the articulatory categories. Unless the speaker is some sort of perfect automaton, the realizations of the binary intentions attributed to him will always exhibit a certain degree of variability.

Perception: The transcriber's task may be very crudely characterized as an attempt to determine what the speaker's intentions were for each articulatory category for each speech sound segment. Because the distributions for a given category overlap, even a 'perfect' transcriber will not be able to give a totally accurate accounting of the speaker's intentions. In fact, his error will be directly related to the amount of overlap. Thus, the 'perfect' transcriber is neither perfect nor a transcriber but an ideal mathematical device which can utilize all the information contained in the distributions produced by the speaker. Henceforth in this paper, the mathematical device will be called an 'Ideal Observer'; a 'live' transcriber who actually makes transcriptions will be called a 'Non-Ideal Observer'.

The following is a description of how a Non-Ideal Observer processes the Signal (Ideal or Non-Ideal) in order to make a transcription.

For each segment, the transcriber receives a complex of sensory data, which will be symbolized X. This complex, X,will consist of information regarding the physical (i.e. acoustic) realization of the speaker's intentions. The information may be expressed in terms of the articulatory categories; therefore, x_{ij} is a single datum where i is a given segment and j is a particular category. Each segment received by the transcriber is then defined as the set:

$$X_i = \{x_{i1}, x_{i2}, \ldots x_{ij}, \ldots x_{i16}\} . \qquad (1)$$

Associated with each segment-category intersection x_{ij} are two probabilities or likelihoods. One is the likelihood that x_{ij} arose from the speaker's intention to give a 'Plus' value to the category j; the other is the likelihood that x_{ij} arose from the speaker's intention to give a 'Minus' value to the category. (Henceforth, only one category will be considered. The process, of course, is repeated for all the categories. 'Zero' intersections have no relevance to the process, since there is no intention attributable to the speaker.)

All the information relevant to the transcriber for describing the speaker's intention at a segment-category intersection may be expressed as a single-number likelihood ratio (that is, the ratio of the two likelihoods discussed above):

$$\lambda(x_{ij}) = \frac{f(x_{ij}|\text{'Plus'})}{f(x_{ij}|\text{'Minus'})} . \qquad (2)$$

If many segments are described in terms of a given category, the likelihood ratio can be plotted on a one-dimensional axis. Any monotonic transformation of likelihood ratio will be equally useful; the natural logarithm of likelihood ratio leads to convenient statistics (Tanner and Birdsall, 1958). $\text{Log}_e\ \lambda(x_{ij})$ is plotted on the abscissa in Fig. 5. The ordinate is the probability density of $\log_e\ \lambda(x_{ij})$. The left-hand distribution in Fig. 5 is conditional upon the 'Minus' intention; the right-hand distribution is conditional

upon the 'Plus' intention. The two distributions are assumed to be normal and to have equal variance.

Insert Figure 5 about here

Fig. 5 preserves intact all the information contained in Fig. 4. If the values of Fig. 5 could be calculated, then the difference in the means of the distributions divided by the standard deviation would yield a detectability index (d') of the speaker's use of category j.

Only the mathematical device (Ideal Observer) can determine the exact value of this d', so it will be subscripted: d'_{IO}. Since a Non-Ideal Observer can only estimate the likelihoods and their ratio, his performance will be somewhat less accurate than the Ideal Observer's. The Non-Ideal Observer's performance has the effect of increasing the variance of the two distributions, thus depressing the value of the detectability index. The d' expressing the Non-Ideal Observer's detection of the speaker's intentions in regard to category j is also subscripted: d'_{NIO}.

In summary, the Non-Ideal Observer operates on a sensory datum, x_{ij}, and the likelihoods of the intentions giving rise to it. His estimates of the likelihood ratio for each x_{ij} will be at variance with the likelihood ratios of an Ideal Observer operating on the same data. The distributions in Fig. 6 express the performance of a Non-Ideal Observer processing the same information found in Figs. 4 and 5.

Insert Figure 6 about here

It is postulated that

$$d'_{NIO} < d'_{IO} \tag{3}$$

Decision: What does the transcriber do with the likelihood ratios he estimates? Certainly he doesn't actually plot their distributions as in Fig. 6. It was claimed that any particular sensory datum, x_{ij}, yields a single-number likelihood ratio $\lambda(x_{ij})$. This value (ignoring the logarithmic transformation for the moment) will range from 0 to $+\infty$.

If $\lambda(x_{ij})$ is very large, it is only reasonable to suppose that the transcriber will decide that the speaker's intention was to make category j 'Plus' for the segment i. Likewise, if $\lambda(x_{ij})$ is very small, the transcriber will decide that the intention was to make category j 'Minus' for segment i. This suggests that the transcriber may adopt a particular value of $\lambda(x_{ij})$, i.e. β, and establish a decision rule such that he will respond

$$\left\{\begin{matrix}\text{'Plus'}\\ \\ \text{'Minus'}\end{matrix}\right\} \quad \text{if } \lambda(x_{ij}) \left\{\begin{matrix}>\\ \\ <\end{matrix}\right\}\beta \tag{4}$$

The probability that the transcriber will respond 'Plus' when the speaker's intention was indeed 'Plus' is the area, to the right of β, under the probability density curve $f(x_{ij}|\text{'Plus'})$. The probability that the transcriber will respond 'Plus' incorrectly, i.e. when the speaker's intention was indeed 'Minus', is the area, to the right of β, under the probability density curve $f(x_{ij}|\text{'Minus'})$. These two probabilities may be symbolized as follows:

$$p(R^+|\text{'Plus'}) = \int_{\beta}^{\infty} f(x_{ij}|\text{'Plus'})\, dx_{ij} \; ; \tag{5}$$

$$p(R^+|\text{'Minus'}) = \int_{\beta}^{\infty} f(x_{ij}|\text{'Minus'})\, dx_{ij} \; , \tag{6}$$

where R^+ means a 'Plus' response by the transcriber.

Fig. 7 is identical to Fig. 6, except that β, $p(R^+|\text{'Plus'})$, and $p(R^+|\text{'Minus'})$ are shown.

Insert Figure 7 about here

If the two distributions are normal and have equal variance, then $p(R^+|\text{'Plus'})$ and $p(R^+|\text{'Minus'})$ describe completely the information contained in Fig. 7. (The other probabilities, $p(R^-|\text{'Plus'})$ and $p(R^-|\text{'Minus'})$, are merely the complements of $p(R^+|\text{'Plus'})$ and $p(R^+|\text{'Minus'})$, respectively.) A table has been constructed (Elliott, 1959) to find the d' value from $p(R^+|\text{'Plus'})$ and $p(R^+|\text{'Minus'})$. The two probabilities can be estimated from the transcriber's actual transcription; this is discussed below.

Independence of d' and β: It will be seen by inspection of Fig. 7 that β can assume any value and not affect d'. This is a crucial aspect of the Theory of Signal Detectability. β is a direct measure of the Observer's response bias due to prior probabilities, affective state, payoffs, etc. In most theories of perception (notably threshold theories), response bias is not accounted for separately from the perceptibility score; it is merely another one of the factors contributing to the overall score. The detectability index d' is, then, a direct measure of an Observer's ability to detect a given Signal, regardless of his response bias.

Estimation of response probabilities: For each category j, a confusion matrix can be constructed from the transcription. In order to construct the matrix, the acutal intentions of the speaker must be known or assumed to be known.

The intentions of the speaker producing an Ideal Signal are assumed to be accurately describable by the speaker himself. This implies that the speaker must be familiar with the system of articulatory categories, and must be able to exert conscious control over his speech parameters. Conceivably, this speaker could produce utterances from some sort of 'script' describing beforehand the articulatory positions and processes to be used.

The intentions of the deviant speaker producing the Non-Ideal Signal are defined as the analphabetic phonetic description, in terms of the Standard, of the target sounds he attempts.

In both cases, the intentions are described as the Plusses and Minuses occurring at segment-category intersections.

The confusion matrix is constructed as follows: The row entries in the matrix are the Plus and Minus values of the speaker's intentions for category j. The column entries are the Plus and Minus values the transcriber indicates in making his transcription (Of course, in making a differential transcription of deviant articulation, the transcriber makes no overt indication of the segment-category intersections in which he thinks the speaker's production of the target sound agrees with the description of the intersection for that target sound in the Standard). In the cells of the confusion matrix are tabulated the frequencies of agreement and disagreement between the actual intentions

(row entries) and the intentions observed by the transcriber. An example of a confusion matrix for a given category is shown in Fig. 8.

Insert Figure 8 about here

The labels in the cells stand for the following: (a) The speaker's intention was 'Plus' and the transcriber detected it as 'Plus'. (b) The speaker's intention was 'Plus' and the transcriber detected it as 'Minus'. (c) The speaker's intention was 'Minus' and the transcriber detected it as 'Plus'. (d) The speaker's intention was 'Minus' and the transcriber detected it as 'Minus'.

The sum $a + b$ is the total number of segments for which the speaker intended a 'Plus' for category j. The sum $c + d$ is the total number of segments for which the speaker intended a 'Minus' for category j. The sum $a + b + c + d$ is the total number of segments (target sounds in the case of the deviant speaker) uttered by the speaker in the sample.

The probabilities given in equations (5) and (6) may be estimated by the following formulas:

$$p(R^+|\text{'Plus'}) = \frac{a}{a + b} \tag{5'}$$

and

$$p(R^+|\text{'Minus'}) = \frac{c}{c + d} \tag{6'}$$

The detectability index (d') for category j can be found by referring these two probabilities to Elliott's (1959) table.

Practical range of d': Theoretically, d' approaches infinity as $p(R^+|\text{'Plus'})$ approaches 1 and $p(R^+|\text{'Minus'})$ approaches 0. This approach to infinity is very gradual, however. In the most readily available table of d' (Elliott, 1959), the highest d' is 4.64, when $p(R^+|\text{'Plus'}) = .99$ and $p(R^+|\text{'Minus'}) = .01$. When $p(R^+|\text{'Plus'})$ and $p(R^+|\text{'Minus'})$ are such that d' would exceed 4.64, it is safe to assign a d' value of 4.90. Unless sample size (total number of segments for a given confusion matrix) is very large (e.g. several thousand), even perfect detection can be assigned a d' value of 4.90.

The difference between $d' = 4.64$ and $d' = 4.90$ is roughly proportional to the difference between any two adjacent d' values in the Elliott table.

Use of d' in the evaluation of 'deviant' articulation

The Non-Ideal Signal: The Non-Ideal Signal can differ from the Ideal Signal in two ways: (1) The means of the two distributions (cf. Fig. 4) may be closer together, and/or (2) The variance of the distributions may be greater. It is therefore postulated that the detectability of the Non-Ideal Signal is less than the detectability of the Ideal Signal:

$$d'_{NIS} < d'_{IS}, \tag{7}$$

given that the Observer is the same in both cases.

Maximum d' for the Non-Ideal Observer: The greatest d' value a 'live' transcriber can achieve for a given category is for an Ideal Signal. This maximum d' can be determined experimentally for any transcriber or group of transcribers.

Efficiency of the deviant speaker: This paper has defined two kinds of Signals and two kinds of Observers. The relationship of these in a diagram analogous to a one-way communication channel (after Tanner and Birdsall, 1958) is shown in Fig. 9.

Insert Figure 9 about here

By referring to the positions of the 'switches' in Fig. 9, the cumbersome subscripts used with the various d' values can be simplified. The two postulates, equations (3) and (7), can be restated as follows:

$$d'_{.2} < d'_{.1} \tag{3'}$$

and

$$d'_{2.} < d'_{1.} \tag{7'}$$

With the same Non-Ideal Observer (transcriber) used for making transcriptions of both Ideal Signals and Non-Ideal Signals, the efficiency of the speaker producing the Non-Ideal Signal can be estimated by

$$\hat{\eta}_j = \frac{d'_{22j}}{d'_{12j}} \tag{8}$$

for each category j. Thus, the transcriber's d' for the deviant speaker (Non-Ideal Signal) is weighted by his d' for the Ideal Signal.

The efficiency score is dependent on both the Non-Ideal Signal and the transcriber. Therefore, transcribers can be used interchangeably only if the efficiency of the Non-Ideal Signal relative to the Ideal Signal (i.e. $\hat{\eta}_j$) is the same (or nearly so) for both transcribers. An equivalent requirement is that the efficiency of the transcribers relative to each other must be the same for both signals. In other words, if

$$\frac{d'_{22j,\ \text{Observer 1}}}{d'_{12j,\ \text{Observer 1}}} = \frac{d'_{22j,\ \text{Observer 2}}}{d'_{12j,\ \text{Observer 2}}}, \tag{9}$$

then

$$\frac{d'_{22j,\ \text{Observer 1}}}{d'_{22j,\ \text{Observer 2}}} = \frac{d'_{12j,\ \text{Observer 1}}}{d'_{12j,\ \text{Observer 2}}}. \tag{10}$$

It should be noted that this requirement is quite different from one which would require that every transcriber's detection index be the same given a particular category of a particular Signal. The requirement stated in equations (9) and (10) means only that the Non-Ideal Observers' relationship to each other be independent of the relationship among the detectabilities (as would be determined by an Ideal Observer) of the Signals involved.

Even if the required equalities of equations (9) and (10) are not met, the transcribers may differ from each other in some regular way. In such a case, the $\hat{\eta}_j$'s of one transcriber can be adjusted by some constant in order to achieve interchangeability. Preliminary data from an ongoing pilot study indicate that the requirements of equations (9) and (10) can be met, with no adjustment or correction factor needed.

Subsets: It may be useful and desirable to estimate $\hat{\eta}$ scores for subsets of target sounds. For instance, it may be interesting to obtain five $\hat{\eta}$'s for the category Voiced-Voiceless: (1) All sounds (in the sample); (2) Vowels only; (3) Consonants only; (4) Stops; (5) Non-Nasal continuants. (In English, it may be found that $\hat{\eta}_{\text{voicing}}$ is very high overall, but that this score is inflated by inclusion of the vowels.) It is

up to the investigator to decide what kinds of subsets to look at, and to justify them on linguistic or practical grounds.

Of course, two d' scores must be obtained for each subset; one for the deviant speaker and one for the Ideal Signal. Thus,

$$\hat{\eta}_{jk} = \frac{d'_{22jk}}{d'_{12jk}} \tag{11}$$

for each subset k of category j.

Discussion

Limitations: The method of transcription and analysis presented in this paper is intended to describe only a very small part of what may be termed deviant speech. The method has value only when there are clearcut segments which may be compared, in one-to-one fashion, with the segments of some standard or target. This limitation excludes from consideration such interesting phenomena as omission, intrusion, metathesis, etc., at least when a quantitative evaluation is needed.

Another limitation is the specificity or 'fineness' of phonetic description achievable by a set of binary categories. The ad hoc addition of categories as they are needed seems to be of doubtful value; eventually the system would be unwieldy and ridiculously complex. The writer feels, however, that the category system as it stands approaches the limit of reliable perceptibility by human observers. Investigations requiring finer discriminations would perhaps be better handled by the various instrumental techniques available.

Applications. The differential transcription method and its statistical analysis are being developed for investigations of the speech of very young children. The method is intended to fill a need for quantitative evaluation of articulation so that a longitudinal plot can be made of a child's development of speech sound specificity (Sharf, Baehr, and Fleming, 1967).

The method should also prove to be a useful tool in other investigations in which speech samples are to be compared, either against each other or against a standard. This would include studies of dialect, speech disorders, etc.

With appropriate modifications, the method may be useful as a clinical tool in both evaluation and therapy planning.

Finally, the method may be of considerable pedagogical use for evaluating and training second-language learners.

Summary

A method for the quantitative analysis of deviant articulation has been proposed. The method is based on (1) analphabetic transcription using binary articulatory categories, and (2) analysis of the transcription in terms of the perceptual performance of the transcriber, as measured by the Theory of Signal Detectability.

Acknowledgments

The writer wishes to thank Donald J. Sharf, Ronald S. Tikofsky, Joan Morley, Michael J. Clark and Ronald Fleming for their most helpful criticism and suggestions.

The sections of this paper dealing with signal detectability would have been impossible without the considerable assistance of Wilson P. Tanner, Jr., Marilyn S. Berman, and Gordon W. Wilcox. The writer thanks these people for their generous contribution of time and knowledge.

References

Clarke, F. R., Birdsall, T. G., and Tanner, W. P., Jr. Two types of ROC curves and definitions of parameters. J. acoust. Soc. Amer., 1959, 31, 629-630.

Elliott, Patricia. Tables of d'. Techn. Rep. No. 97, Univer. Michigan: Electronic Defense Group, 1959.

Halle, M. On the bases of phonology. In J. Fodor & J. Katz (Eds.), The structure of language: readings in the philosophy of language. Englewood Cliffs, N. J.: Prentice-Hall, 1964, Pp. 324-335.

Jakobson, R., Fant, C. G. M., and Halle, M. Preliminaries to speech analysis (4th Ed.) Cambridge, Mass.: M.I.T. Press, 1963.

Jakobson, R., and Halle, M. Fundamentals of language. 's-Gravenhage: Mouton, 1956.

Pike, K. L. Phonetics: a critical analysis of phonetic theory and a technic for the practical description of sounds. Ann Arbor: Univer. Michigan Press, 1943.

Sharf, D. J., Baehr, T. J., and Fleming, Katherine. A system for the analysis of speech sound development. Techn. Rep. 14, USPHS Grant HD 01368, Center Human Growth and Dev., Univer. Michigan, 1967.

Swets, J. A. (Ed.) Signal detection and recognition by human observers. New York: Wiley, 1964.

Swets, J. A., Tanner, W. P., Jr., and Birdsall, T. G. Decision processes in perception. Psychol. Rev., 1961, 68, 301-340.

Tanner, W. P., Jr., and Birdsall, T. G. Definitions of d' and η as psychophysical measures. J. acoust. Soc. Amer., 1958, 30, 922-928.

Table 1

Standard for the writer's variety of Midwest American

| | p^{-} | p^{+} | t^{-} | t^{+} | c^{-} | c^{+} | k^{-} | k^{+} | $\check{D}$ | b^{-} | β^{-} | d^{-} | d° | $ʤ^{-}$ | $ʤ^{\circ}$ | g^{-} | g° | f^{-} | f° | s^{-} | s° | $ʃ^{-}$ | $ʃ^{\circ}$ | v^{-} | v° | z^{-} | z° | $ʒ^{-}$ | $ʒ^{\circ}$ | θ^{-} | θ° | $ð^{-}$ | $ð^{\circ}$ | $ɾ^{-}$ | $ɾ^{\circ}$ | $ʧ^{-}$ | $ʧ^{\circ}$ | $ʤ^{-}$ | $ʤ^{\circ}$ | m^{-} | m° | n^{-} | n° | $ŋ^{-}$ | l^{-} | $l^{\circ 1}$ | $l^{\circ 2}$ | $ɬ$ | i | ɪ | I | $ʊ^{-}$ | $ʊ^{\circ}$ | e | ɛ | æ | ɑ | A | ɔ | O | o | u | ʊ | ⊔ | U | ʊ | ə | ɚ | R | h |
|---|---|---

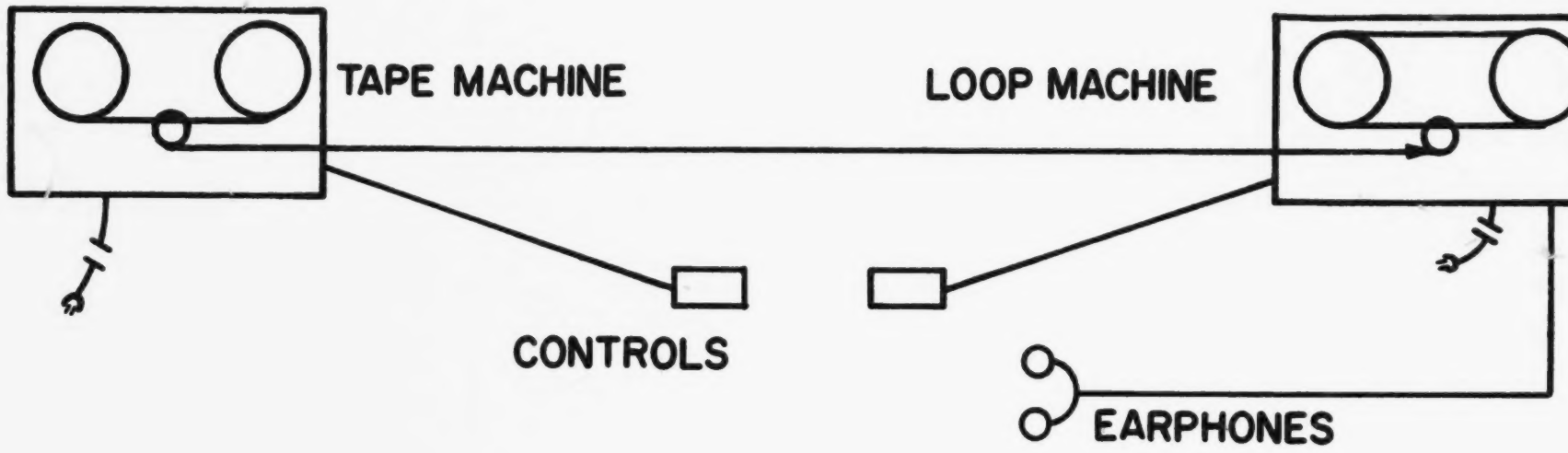

Fig. 1. Schematic of transcription equipment.

	f°			ə			n̄			ɛ			Ď			I			c̄	
	+	-		+	-		+	-		+	-		+	-		+	-		+	-
VOC		X	VOC	X		VOC		X	VOC	X		VOC	X		VOC	X		VOC		X
CON	X		CON		X	CON	X		CON		X	CON	X		CON		X	CON	X	
INT		X	INT			INT		X	INT			INT	X		INT			INT	X	
EDG	X		EDG			EDG		X	EDG			EDG		X	EDG			EDG		X
P-1	X		P-1		X	P-1		X	P-1		X	P-1		X	P-1		X	P-1	X	
P-2		X	P-2		X	P-2		X	P-2		X	P-2		X	P-2		X	P-2		X
M-1			M-1		X	M-1			M-1	X		M-1			M-1	X		M-1		
M-2		X	M-2			M-2		X	M-2			M-2		X	M-2			M-2		X
CLO	X		CLO		X	CLO	X		CLO		X	CLO	X		CLO	X		CLO		X
OPN			OPN		X	OPN			OPN		X	OPN			OPN	X		OPN		
NAS		X	NAS		X	NAS	X		NAS		X	NAS		X	NAS		X	NAS		X
VCD		X	VCD	X		VCD	X		VCD	X		VCD	X		VCD	X		VCD		X
TNS			TNS		X	TNS		X	TNS		X	TNS		X	TNS		X	TNS		X
EGR	X		EGR	X		EGR	X		EGR	X		EGR	X		EGR	X		EGR	X	
PLM	X		PLM			PLM	X		PLM			PLM	X		PLM			PLM	X	
GLT			GLT			GLT			GLT			GLT			GLT			GLT		

Fig. 2. Transcription form with transcription of the word 'phonetic'.

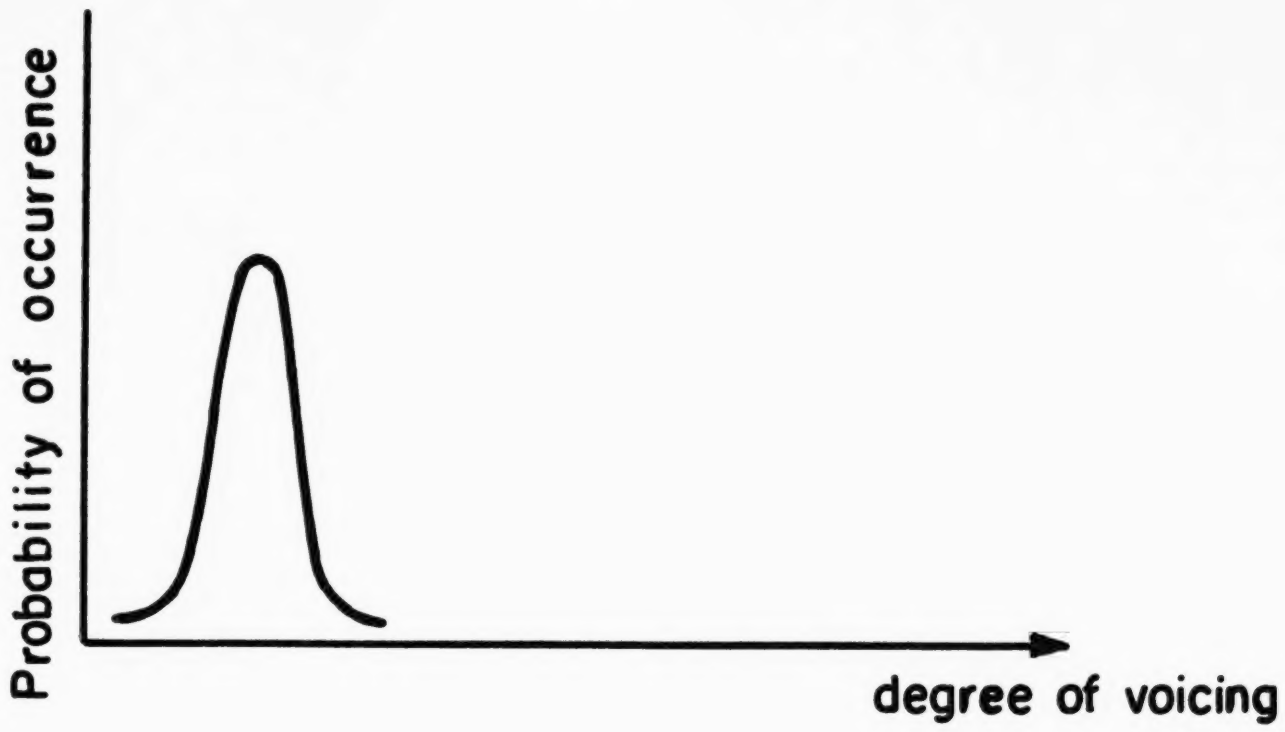

Fig. 3. Hypothetical distribution of voiceless sounds.

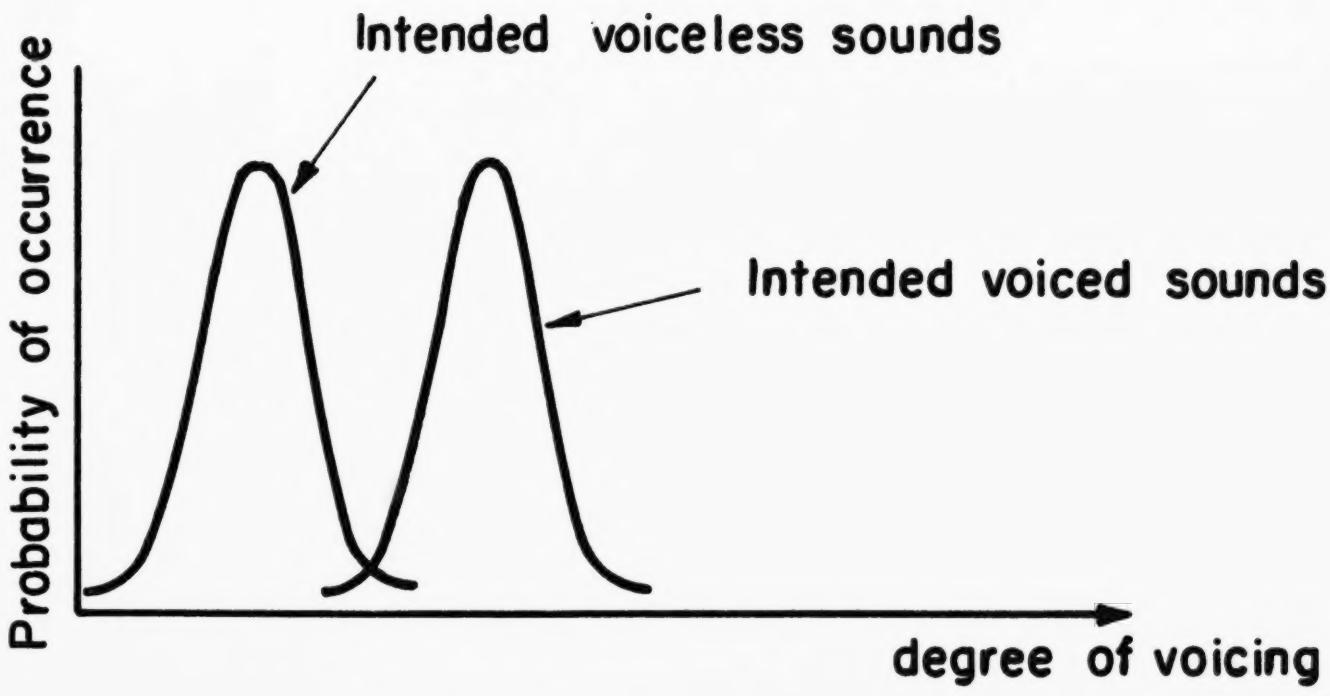

Fig. 4. Hypothetical distribution of voiced and voiceless sounds.

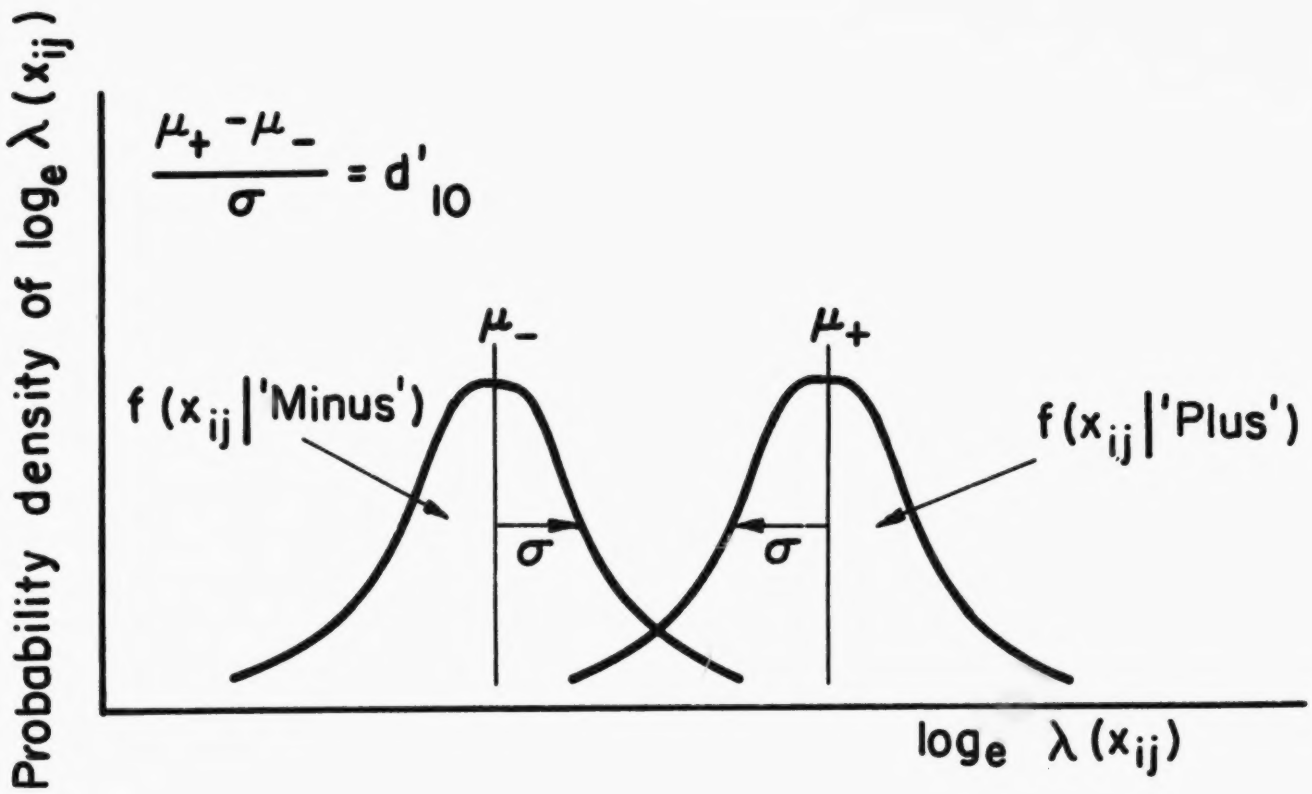

Fig. 5. Ideal signal for category j as detected by ideal observer.

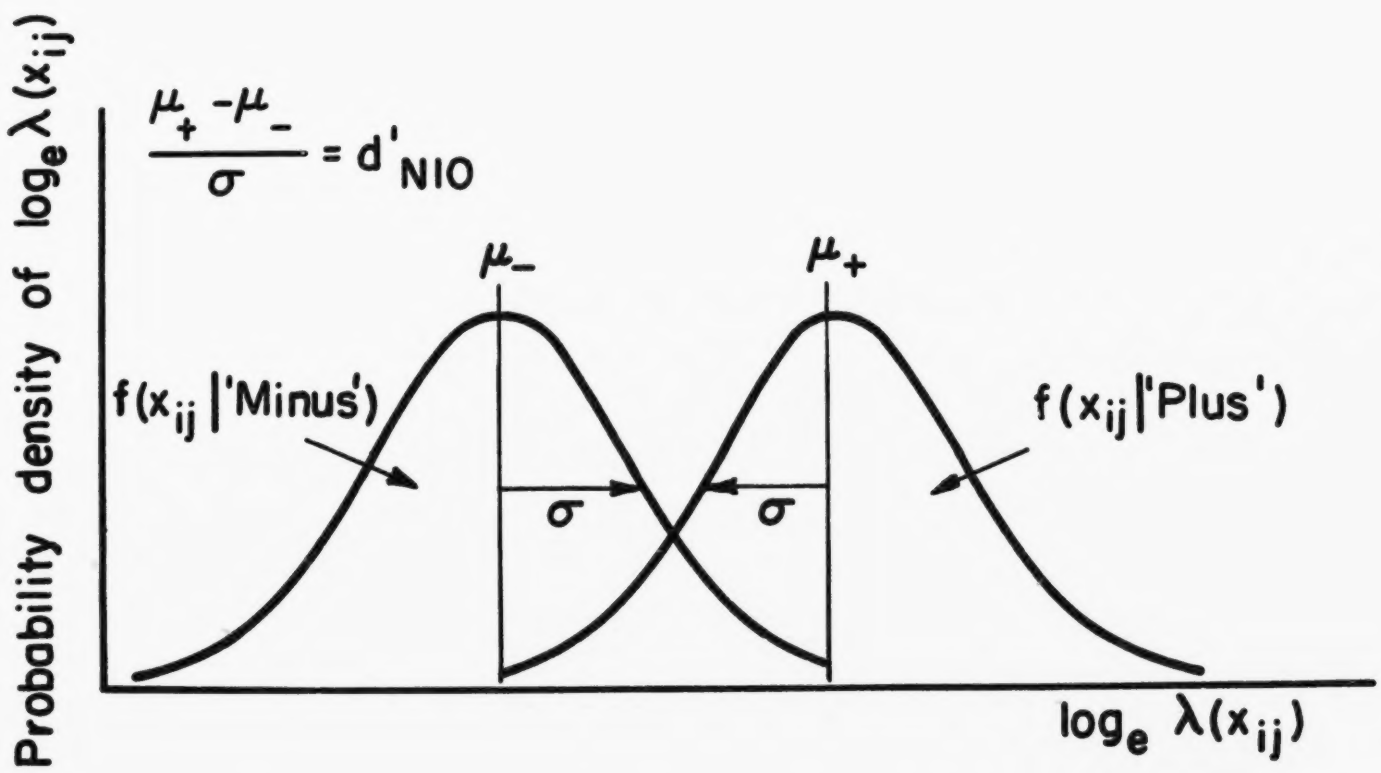

Fig. 6. Ideal signal for category j as detected by non-ideal observer.

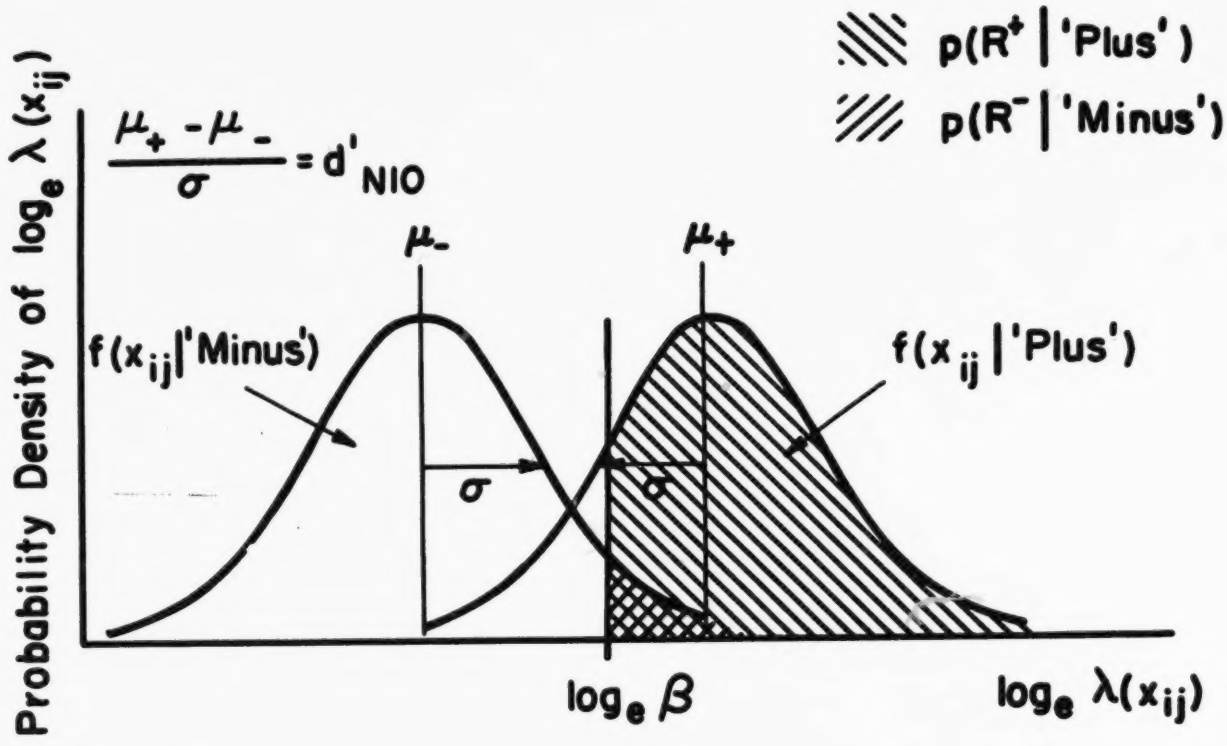

Fig. 7. Indication of β and response probabilities for Fig. 6.

Transcriber's Responses

Speaker's Intentions	R^+	R^-
'Plus'	a	b
'Minus'	c	d

Fig. 8. Confusion matrix summarizing transcriber's responses in regard to speaker's intentions for a given category.